Predicting Helical Hairpins from Sequences by Monte Carlo Simulations

PHILIPPE DERREUMAUX*

Laboratoire de Biochimie Théorique, UPR CNRS 9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France

Received 18 October 1999; accepted 10 January 2000

ABSTRACT: The key problem in polypeptide-structure prediction is with regard to thermodynamics. Two factors limit prediction in *ab initio* computer simulations. First, the thermodynamically dominant conformations must be found from an extremely large number of possible conformations. Second, these low-energy forms must deviate little from the experimental structures. Here, we report on the application of the diffusion-controlled Monte Carlo approach to predict four α-helical hairpins with 34–38 residues by global optimization, using an energy optimized on other supersecondary structures. A total of seven simulations is carried out for each protein starting from fully extended conformations. Three proteins are correctly folded (within 3.0 Å rms from the experimental structures), but the fourth protein cannot distinguish between several equienergetic conformations. Possible improvement of the energy model is suggested. © 2000 John Wiley & Sons, Inc. J Comput Chem 21: 582–589, 2000

Keywords: Monte Carlo simulations; optimized potential; *ab initio* structure prediction; α-helical hairpins; global optimization

Introduction

Determining the unfolding–folding mechanism of proteins and predicting the native structure of proteins from their amino acid sequences are two interrelated aspects of the protein folding problem. The experimental and theoretical backgrounds for protein folding started in the early 1970s with the pioneering work of Anfinsen,[1] Levitt and Warshel,[2] and Taketomi et al.[3] Since then, our knowledge of both protein thermodynamics and kinetics has increased enormously through the combination of ingenious experimental studies,[4] protein engineering experiments,[5] Monte Carlo (MC) simulations, and analytical approaches using simple protein models.[6] However, *ab initio* structure prediction by global optimization of an energy function is still one step behind.[7–9] To overcome this problem, new methods have emerged such as the threading approach, comparative modeling, and biased simulations using secondary structure predictions

Correspondence to: P. Derreumaux; e-mail: philippe@igs.cnrs-mrs.fr

*Also at Laboratoire Information Génétique & Structurale, CNRS-UMR 1889, Marseille, France

and multiple sequence alignments. Both comparative modeling and threading approaches are aimed at building a structural model of a protein on the basis of close sequence similarity to a template protein of known structure. An excellent review of these methods was given by Koehl and Levitt.[10] Although good accuracy in the prediction is often achieved for query proteins sharing more than 30% identity with their templates, all these methods still rely on an energy function to fold (simulations) or to discriminate the native structure from a set of unfolded structures (threading approach and comparative modeling).

One way to design an effective energy model is to study short peptides. These have only marginally stable conformations in aqueous solution, but their energy hypersurfaces can be explored extensively for chain lengths up to around 25 residues by molecular dynamics (MD) simulations,[11] Monte Carlo (MC) and MD simulations in the multicanonical ensemble,[12] and a variety of MC-based[9, 13] and constraint-based conformational approaches.[14] In this study, we extend our work on the combined use of the diffusion-controlled process Monte Carlo method[15] and an optimized potential for polypeptides (OPEP).[16] This method is used to predict conformational ensemble averages for helical hairpin structures. It uses, on average, six beads per resldue (i.e., N, H, Cα, C, O, and one particle for the side chains) and builds a structure for the target sequence by optimization of an effective potential without using experimental data.

α-Helical hairpin (AHH) models with 34–38 residues, composed of two helices connected by a turn, have been chosen because they occur frequently in proteins, they cannot be handled by the previously described methods and because there are experimental data available on the structures and stabilities of the entire molecules and their two helical constituents in isolation. Thus, the short-range (between residues i and $\leq i + 4$) versus long-range (between distant residues i and $> i + 5$) effects can be investigated and an effective energy model allowing discrimination of the native structure can be tested or elaborated. Furthermore, as a recent study has shown, agreement between the OPEP-MC simulations and nuclear magnetic resonance (NMR) spectroscopy was presented for 23 polypeptides including a de novo-designed AHH peptide stabilized by a disulfide bond (ALIN), but structural deviation was found in the helix lengths for another AHH peptide, termed ATA.[16] The first and second helix lengths were predicted to be 10 and 5 residues versus 17 and 13 residues by NMR.

We also sought to determine if such disagreement for ATA is accidental or results from an inherent problem in the energy model. To this end, the ALIN and ATA structure ensembles are examined and the helix content of the two peptides spanning the ATA peptide is calculated by the present approach and compared with the percentage determined from the circular dichroism (CD) signal. We then give the details of the results of the structure prediction for two other AHH polypeptides, comparing the lowest energy structures and the experimental data. Finally, the implications of our results for structure prediction are discussed.

Materials and Methods

As a compromise between structural resolution, energetics, and search efficiency, a simplified off-lattice representation is used. The main chain atoms are treated explicitly and the side chain atoms of all residues except proline are represented by a bead placed at an appropriate position with an appropriate van der Waals radius. All internal coordinates are free to vary except the peptide bond dihedral angles, ω, which are fixed at 180°. Such a model reproduces experimental structures exactly.

An essential aspect of the energy model is the representation of aqueous solution effects. The analytic form of the OPEP1.0 function and its set of parameters have been optimized on the structures of four polypeptides with 10–28 residues in aqueous solution, by maximizing the energy gaps between the native states and representative ensembles of low-energy unfolded states. This function was then tested on 20 polypeptides, which were not themselves used during the optimization procedure. It is expressed by:

$$E = w_H E_{HB1} + w_{HH} E_{HB2} + w_L E_L + w_{SC} E_{SC,SC} + w_A E_{A,A} + \sum_{20} w_P^{\alpha} E_P^{\alpha} + \sum_{20} w_P^{\beta} E_P^{\beta} + w_P E_P \quad (1)$$

This potential goes beyond the simple pairwise potential between side chains, $E_{SC,SC}$, which is represented by a 12-6 potential if the interaction is hydrophobic in character and by a 6-potential for hydrophilic interactions, considering all 20 natural amino acids. The E_{HB1} (two-body interaction) and E_{HB2} (four-body interaction) terms are the two components of the main-chain H-bond potential. The E_L harmonic term maintains all internal coordinates (except ϕ and ψ) near their equilibrium values. The $E_{A,A}$ terms are pairwise contact potentials between α-carbons. The E_P^{α} and E_P^{β} terms are the α-helix and

β-sheet propensity potentials for the amino acids, respectively. The E_P component is a penalty term applied to the residues preceding proline in the sequence if their ϕ values are $\geq 0°$. The values of the weights, w, and details on the components are given in Derreumaux.[16]

Because all test models presented here correspond to α-helical proteins with or without one disulfide bond, we examine the role of the E_P^α values and the disulfide bond energy term on the calculated energy surfaces. First, the α-helix propensity potentials do not strongly favor the formation of the helices. Rather, residues alanine (−0.3 kcal/mol), arginine, and leucine (−0.15 kcal/mol) favor α-helix formation, glutamic acid is neutral (0.0 kcal/mol), but all others oppose helix formation ($\geq$+0.7 kcal/mol). Furthermore, the value of $w_P^\alpha E_P^\alpha$ for Ala is 10% of the value of $w_{SC}E_{SC,SC}$ between two leucine residues. Second, the disulfide bond interaction is not encoded in the input data, but is treated as a pairwise contact potential. Because the energy assigned to this contact is only twice that between two leucine residues, the disulfide bond energy term is not essential in stabilizing a conformation. In summary, the OPEP potential is not biased toward the formation of a disulfide bond and helices. The predicted energy landscape for any single protein results from a subtle balance between backbone energy terms (including the α-helix and β-strand preferences and H bonds) and long-range residue contact terms.

To find the lowest energy state of the simplified chain on the energy hypersurface, we use the diffusion process-controlled MC method.[15, 16] This method is not guaranteed to generate a Boltzmann-distributed (canonical) ensemble of conformations at temperature T, but was found to locate the lowest energy region of 24 polypeptide chains with 7–38 residues independently of the complexity of the native fold.[16] The concept behind this MC method is to search trial conformations that would be thermodynamically and kinetically accessible from a given conformation in a reasonable time. Using diffusion considerations of the main-chain atoms in ϕ, ψ space, this kinetic requirement is translated into a condition on the maximal angular deviations that all residue can undergo at each MC step and thus limits the number of residues to be moved in the next conformation. (The system is moved from its old conformation to a new conformation by forcing the residues to adopt their new ϕ, ψ values in the restricted volume of the old conformation.) The next conformation is accepted or rejected by the Metropolis criterion.[17]

Results

The four polypeptides, which are the subject of prediction, have been characterized by NMR spectroscopy as adopting the AHH motif in aqueous solution. ALIN[18] and ATA[19] are two de novo-designed proteins, Z38 is a variant of the first two α-helices of the 59-residue B-domain of protein A, and Z34C is a 34-residue disulfide-bonded variant of Z38.[20] The sequences of these peptides and of the two peptides spanning the helices in the ATA peptide, namely A1 and A2, are shown in what follows (Ac: acetyl; Suc: succinyl; abbreviations for the amino acids can be found in ref. 16):

ALIN	Ac—YERDELMACL KKATNGTSGN TEREDLIACV KRATH—NH_2
ATA	Suc—DWLKARVEQE LQALEARGTD SNAELRAMEA KLKAEIQK—NH_2
Z38	AVAQSFNMQQ QRRFYEALHD PNLNEENRNA KIKSIRDD
Z34C	FNMQCQRRFY EALHDPNLNE EQRNAKIKSI RDDC
A1	Suc—DWLKARVEQE LQALEARG—NH_2
A2	SNAELRAMEA KLKAEIQK—NH_2

Table I reports the calculated helix contents for the four proteins using standard secondary structure prediction methods. The goal of these methods is to predict whether the residue at the center of a segment of typically 13–21 adjacent residues is in a helix, strand, or in nonregular secondary structure. The combined use of eight methods points to a reasonable agreement between the consensus predicted and experimental percentage helicities. These data are not included in the OPEP-MC simulations.

For each AHH peptide and constituent peptide of ATA, five simulations at 600 K and two simulations at 800 K were performed for 6000 MC steps, starting from fully extended conformations. The converged conformers did not change with temperature. We analyzed the generated ensemble averages by the following quantities: (1) root-mean-square deviations (rmsd's) from the NMR structures for the main-chain atoms; and (2) lengths of individual helices or percentage helicities. Helix content was computed from the Boltzmann distribution of the number of $i \rightarrow i+4$ hydrogen bonds, provided that ϕ and ψ angles deviate little ($\pm 30°$) from the canonical values of $-60°$ and $-40°$, respectively.

In an earlier study we showed that the OPEP-MC simulations at 500 K reproduced the main structural NMR features of ALIN.[16] (The NMR-derived

TABLE I.
Percentage Helicities of Four Protein Sequences Determined by Eight Standard Secondary Structure Prediction Methods.[a]

Model	Consensus Value[b]	Extreme Values	Experimental
ALIN	60%	40% (GOR4) to 71% (PHD)	41%
ATA	74%	60% (DSC) to 87% (PHD)	79%
Z38	60%	37% (SOPM) to 74% (DSC)	66%
Z34C	62%	27% (GOR4) to 71% (SIMPA96)	76%

[a] The eight methods, namely DPM, DSC, GOR4, HNNC, PHD, PREDATOR, SIMPA96, and SOPM, are available on the world-wide web (http://pbil.ibcp.fr).

[b] The consensus prediction is described in ref. 21.

coordinates have not been reported.) The new simulations at higher temperatures give identical results, suggesting that the lowest energy region has been located. All lowest energy structures have the correct disulfide bond interaction between residues 19 and 29, have two helices spanning residues 6–15 and 23–34 (versus 6–11 and 24–31 experimentally), and satisfy the three interhelical NOEs (nuclear Overhauser effect), namely, between the side chains Leu10–Cys29, Leu10–Val30, and Ala13–Leu26. However, we also find, 2 kcal/mol above the previous structure, a slightly distorted structure with two helices spanning residues 8–15 and 23–33 with a flexible N-terminal segment. Such motion at residues 1–5 is consistent with the fluctuations observed by NMR spectroscopy. [18]

ATA was characterized by NMR spectroscopy to have a helix content of 89% with both termini fluctuating at pH 3.6 and 25°C.[19] The helix I includes residues Asp1–Arg17 and helix II consists of residues Ala23–Glu35. All OPEP1.0 MC simulations generate partly unfolded structures (rmsd ~5 Å for all main-chain atoms and 3 Å for residues 5–33) with 40% helix content. These structures are more stable than the NMR-derived structure by ~3 kcal/mol. Figure 1 shows the lowest energy structure and the NMR structure. In all predicted structures, residues 8–17 or 5–17 are helical (vs. 1–17 in the NMR structure) and residues 25–29 are either helical or random coil (vs. 23–35 in the NMR structure). Both NMR and predicted loop regions encompass residues Gly18–Asn22 and involve a mixture of turn types.

To determine the factors responsible for such a decrease in helicity, helix content was calculated for peptides A1 and A2 in aqueous solution. The predicted helix content averages 45% in A1 (helical from residues Glu8–Ala16) and 34% in A2 (helical from residues Met8–Ala14), very close to the CD results of 34–43% for both peptides[22] and the percentage using the AGADIR algorithm (33% for A1 and 52% for A2).[23] We note that A2 also has an important propensity to form a unique helical segment from residues Asn2–Ala7, 1.5 kcal/mol above the lowest energy state. No medium-range NOEs are available for validating the positions of the predicted helices in A1 and A2.

Why then is the helicity of A1 essentially unchanged and that of A2 substantially decreased in the ATA-predicted conformations? If helicities of both A1 and A2 were affected, the interhelical in-

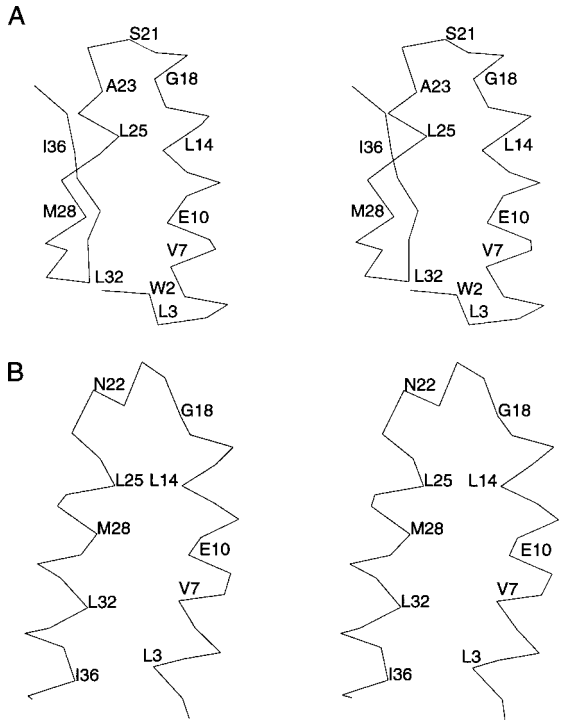

FIGURE 1. Stereoviews of Cα traces for the lowest energy predicted structure (a) and the NMR structure (b) of ATA.

teractions would be overestimated. Because only the helix content of A2 is affected, we examine the structural features of the ATA conformations. In all structures, we see that residues 33–38 are folded against residues 25–29 due to the existence of a γ-turn ($\phi = 68°$, $\psi = -67°$) conformation at Leu32, and make hydrophobic contacts with the A1 segment (e.g., between Leu14 and Leu36). We also find the extremity of A1 compact due to the presence of γ turns, which are generally disallowed for all residues except glycine and asparagine. These findings imply that the local stiffness of the chain and thus the stereochemical properties of the amino acid residues in our potential are not totally consistent with the results from the conformational analysis of residues in high-resolution structures. An important question is whether further evidence for energetic limitations is provided by the sequences discussed in what follows, which were not used during the optimization of the OPEP model.

Z38 has been shown to adopt two helices in an antiparallel arrangement by NMR spectroscopy at pH 7.5 and 8°C.[20] The native conformation is disordered from Ala1 to Ser5 and well-defined between Phe6 and Arg36, with two helices spanning residues 6–18 and 25–36. However, the temperature dependence of the CD spectrum reveals that, even at 8°C, Z38 is not stable. The unfolded conformations have not been characterized experimentally.[20] Application of the OPEP-MC procedure results in an ensemble of structurally distinct conformations of similar energy, deviating by 5 kcal/mol at most. All the structures belong to the α-helical hairpin class, but deviate from the NMR-folded structure by the lengths of the two helices and the list of interhelical NOEs that are satisfied (Table II). Figure 2 shows the lowest energy structure (S1), the second lowest energy structure (S2), and the highest energy structure (S7) superposed on the NMR structure. We see that, of the two helices, helix I is very unstable. It spans residues 6–18 in NMR, 11–19 in S1, 12–17 in S2, and 3–19 in S7.

Based on the marginal stability of Z38, Starovaski et al. designed Z34C by essentially removing the first five unstructured N-terminal residues of Z38 and introducing a disulfide link between residues 5 and 35.[20] This modified sequence was found by NMR spectroscopy to adopt a stable AHH structure, composed of two helices, helix I (residues 3–13) and helix II (residues 20–34), connected by a loop between residues 14–19.[20] The predicted structure ensemble is shown in Figure 3. Two simulations (at 600 and 800 K) generate NMR-like structures as evident from the rmsd values of 3.9 Å for residues 5–34 and the occurrence of helices from residues 3–11 and 25–34 (Fig. 3A). The five remaining simulations generate structures deviating from experiment in two respects: (1) Helix II is formed, the disulfide bond interaction is formed, but helix I is destabilized. Rather, we find a coil structure with a helical segment from residues 10 to 14 folded against residues 3–5 due to the presence of a γ-turn at Arg7 (Fig. 3B). This structure with a rmsd of 4.2 Å has an energy equal to the NMR-like structure just described. (2) Helix II is correctly predicted, but the disulfide bond interaction is not formed and helix I is replaced by two helices from residues 7–14 and 17–20 (Fig. 3C). A γ-turn conformation is also found at Gln4. This conformation with a rmsd of 4.4 Å is 2.4 kcal/mol above the lowest energy structures.

Conclusions

We have presented the application of the OPEP-MC method to predict the lowest energy conformations for four polypeptides 34–38 amino acids in length adopting helical hairpins in solution. Starting from fully extended conformations, the OPEP-MC procedure generates native-like or partially unfolded structures deviating by 3-Å rms from the

TABLE II.
Structural Deviations between the Predicted and NMR-Folded Structures of Z38.

Structures	$rmsd_{10-36}$	$rmsd_{6-36}$[a]	Helix I	Helix II	NOE F14–I35	NOE A17–I32
NMR	0.0	0.0	6–18	25–36	Yes	Yes
S1	3.7	3.2	11–19	22–25, 29–37	Yes	No
S2	4.0	4.3	1–5, 12–17	25–38	No	No
S7	4.0	2.8	3–19[b]	23–34	No	No

[a] rmsd values (in angstroms) are between residues 6 and 36 because residues 1–5, 37, and 38 are disordered experimentally.[20]
[b] Helix is kinked at Gln9–Gln10.

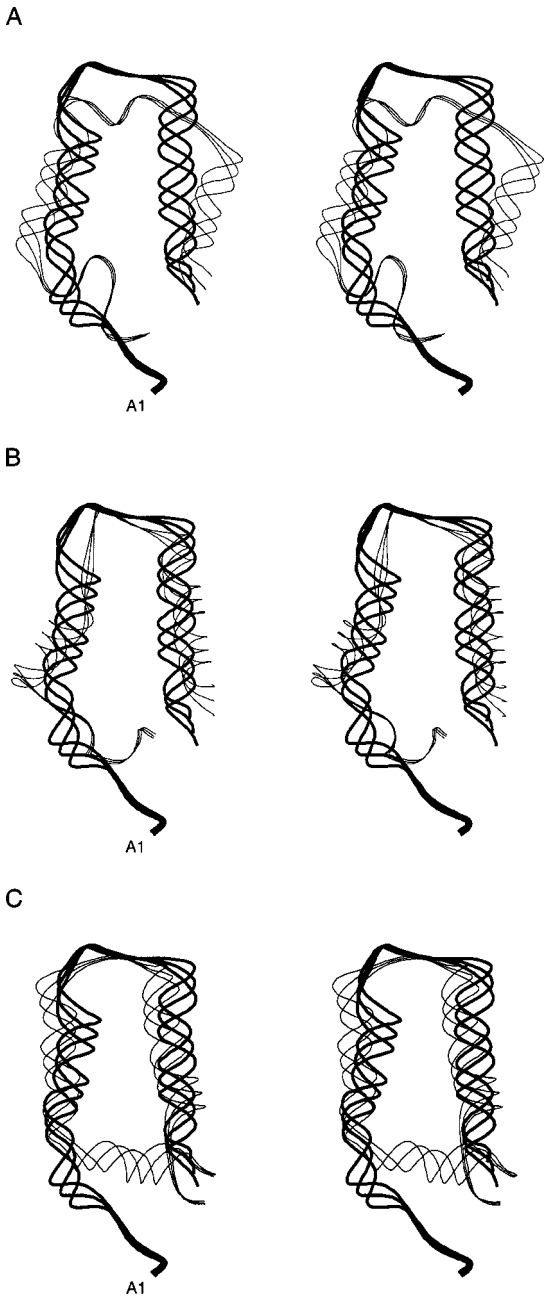

FIGURE 2. Superposition of the ribbon drawings of the lowest (a), the second lowest (b), and the highest (c) energy structures of Z38, on its NMR structure (bold).

NMR structures for the ALIN, ATA, and Z38 proteins, but cannot find the NMR structure among unfolded structures for Z34C. Failure to find this conformation can be assigned to a problem of sampling and/or to limitations of the energy function. Although convergence to the ground state can be slow if the energy gap between the native and low-energy misfolded conformations is not optimal, and

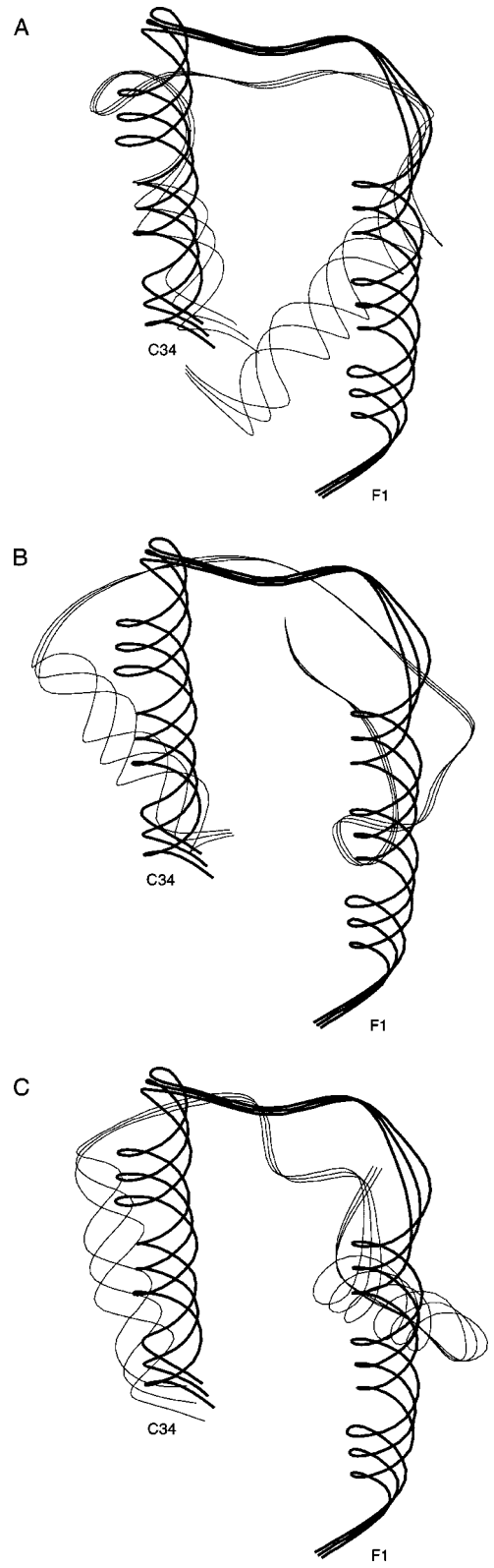

FIGURE 3. Superposition of the predicted Z34C conformations, superposed on the NMR (bold) structure: (a) native-like; (b) partially unfolded; and (c) misfolded conformations.

inadequate sampling cannot be excluded for all sequences and runs within an estimated number of MC steps function of chain length,[16] two factors indicate that a metastable conformation has not been located in these simulations. First, the predicted lowest energy structures of Z34C are more stable than the NMR-derived structure by ~5 kcal/mol. Second, starting from the NMR-derived structure, five runs at 800 K converge to the vicinity of the expected (predicted) conformational ensemble. This result is significant because it shows that simulations starting from conformations having percentage helicities consistent with secondary structure prediction methods would not change the results.

The results presented here and elsewhere[16] are encouraging because the dominant NMR fold is found among the lowest energy structures and the NMR protection factors for the slowest exchanging protons in these molecules are more in line with molten globule forms or folding intermediates of proteins.[22] However, reexamining the local forces will be an obligatory step to make OPEP a more robust discriminating potential. The approximate calculated density of conformational states does not show a large energy gap between native-like and misfolded conformations for the ATA and Z34C models. Analysis of the structural features of the low-energy conformations for the four proteins shows that the short-range interactions controlling the stereochemical properties of the amino acid residues are not always compatible with the expected (ϕ, ψ) maps. Residues arginine, leucine, and glutamine can occupy the region $\phi > 0^\circ$ in the predicted conformations, whereas this is generally disallowed in the equilibrium conformations found within the protein data bank (PDB renamed RSCB).

Determining the correct energy model is no simple task for computer simulations because a compromise between energetics and sampling efficiency must be found, the misfolded conformations themselves depend on the energy function, and the analytic form of the energy function must be tested on a large set of models. For instance, five simulations on Z34C, without the α-helix propensity potential for the amino acids, find NMR-like structures with two helices spanning residues 4–13 and 21–34 (vs. 3–13 and 20–34 in the NMR structure), but five simulations on Pep28[16] do not find the native $\beta\beta\alpha$ state.

Because various terms can be added to the present energy model and the missing factors may vary from one topology to another, it is necessary to examine their influence on representative models such as helical hairpins, three antiparallel β-strands,[24] and nonstandard α-topologies exhibited by the 43-residue protein extracted from pyruvate dehydrogenase.[25] These factors include position-dependent effects on amino acid structural propensities, torsional potentials,[26] $i \rightarrow i + 3$ hydrogen bonds, and N-cap and C-cap effects (i.e., hydrogen bonds between the side chain and the main chain at the N- and C-termini of helices). Such a systematic study will have implications for structure prediction, but also for protein folding, and notably the stages of structural reorganization during the early steps of folding.

Acknowledgments

The author is grateful to Richard Lavery and the anonymous referees for helpful comments on the manuscript.

References

1. Anfinsen, C. B. Science 1973, 181, 223–230.
2. Levitt, M.; Warshel, A. Nature 1975, 253, 694–698.
3. Taketomi, H.; Ueda, Y.; Go, N. Int J Pept Prot Res 1975, 7, 445–459.
4. (a) Lecomte, J. T. J.; Falzone, C. J. Nature Struct Biol 1999, 6, 605–608; (b) Sabelko, J.; Ervin, J.; Gruebele, M. Proc Natl Acad Sci USA 1999, 96, 6031–6036.
5. (a) Tuchscherer, G.; Scheiber, L.; Dumy, P.; Mutter, M. Biopolymers (Pept Sci) 1998, 47, 63–73; (b) Walsh, S. T. R., et al., Proc Natl Acad Sci USA 1999, 96, 5486–5491.
6. (a) Dinner, A. R.; So, S.-S.; Karplus, M. Prot Struct Funct Genet 1998, 33, 177–203; (b) Alm, E.; Baker, D. Curr Opin Struct Biol 1999, 9, 189–196.
7. Mirny, L.; Shakhnovich, E. I. J Mol Biol 1998, 283, 507–526.
8. (a) Skolnick, J.; Kolinski, A.; Ortiz, A. R. J Mol Biol 1997, 265, 217–241; (b) Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. Proc Natl Acad Sci USA 1998, 95, 2932–2937.
9. (a) Pedersen, J. T.; Moult, J. J Mol Biol 1997, 269, 240–259; (b) Lee, J.; Liwo, A.; Scheraga, H. A. Proc Natl Acad Sci USA 1999, 96, 2025–2030.
10. Koehl, P.; Levitt, M. Nature Struct Biol 1999, 6, 108–111.
11. Duan, Y.; Kollman, P. A. Science 1998, 282, 740–744.
12. (a) Schaefer, M.; Bartels, C.; Karplus, M. J Mol Biol 1998, 284, 835–848; (b) Hansmann, U. H. E.; Okamoto, Y. Curr Opin Struct Biol 1999, 9, 177–183.
13. Hoffmann, D.; Knapp, E.-W. J Phys Chem B 1997, 101, 6734–6740.
14. Ishikawa, K.; Yue, K.; Dill, K. A. Prot Sci 1999, 8, 716–721.
15. Derreumaux, P. J Chem Phys 1998, 109, 1567–1574.
16. Derreumaux, P. J Chem Phys 1999, 111, 2301–2310.
17. Metropolis, N. S., et al., J Chem Phys 1953, 21, 1087–1092.
18. Kuroda, Y.; Nakai, T.; Ohkubo, T. J Mol Biol 1994, 236, 862–868.
19. Fezoui, Y.; Connolly, P. J.; Osterhout, J. J. Prot Sci 1997, 6, 1869–1877.

20. Starovaski, M. A.; Braisted, A. C.; Wells, J. A. Proc Natl Acad Sci USA 1997, 94, 10080–10085.
21. Deleage, G.; Blanchet, C.; Geourjon, C. Biochimie 1997, 79, 681–686.
22. Fezoui, Y.; Braswell, E. H.; Xian, W.; Osterhout, J. J. Biochemistry 1999, 38, 2796–2804.
23. Lacroix, E.; Viguera, A. R.; Serrano, L. J Mol Biol 1998, 284, 173–191.
24. De Alba, E.; Santoro, J.; Rico, M.; Jimenez, M. A. Prot Sci 1999, 8, 854–865.
25. Kalia, Y. N., et al., J Mol Biol 1993, 230, 323–341.
26. DeWitte, R. S.; Shakhnovich, E. I. Prot Sci 1994, 3, 1570–1581.